Statistics and Its Interface Volume 4 (2011) 267–272

Real data examples in statistical methods papers: Tremendously valuable, and also tremendously misvalued

K. Y. Williams, Yun Joo Yoo, Amit Patki and David B. Allison*

When a statistical methods paper is submitted to a journal for publication, examples in which the method is applied to real data are highly encouraged by many journals and in some cases are explicitly demanded. In this commentary, we argue that real data examples serve several useful purposes. However, we also argue that in many cases, particularly in the fields of genetics and genomics, there is an implicit or explicit expectation for examples to support purposes for which they are ill-suited and furthermore that these inappropriate expectations have negative consequences for the field. We conclude by noting that real data examples can be tremendously valuable and should continue to be used where appropriate, but that the demands for, expectations of, and conclusions drawn from them need to be scaled back.

Keywords and phrases: Examples, Simulation, Methodology, Statistics, Genomics, Pedagogy, Publishing.

Statistical methods are vitally important in the biological sciences and continue to evolve. This is nowhere more true than in genetic and genomic research. In our experience, reviewers and editors of journals that publish papers offering new statistical methods in genetics and genomics are favorably disposed to papers including a real data example that illustrates the application of the method or methods under study. We believe that the inclusion of real data examples is highly desirable for reasons that we describe below. However, there also seems to be a prevailing belief on the part of many reviewers and editors, especially of high-impact journals, that (a) a real data example is essential, (b) the example should reveal an exciting biological finding, and (c) the method that obtains this exciting finding offers a proof of principle or validation of the method. In contrast, we believe that each component of this tripartite belief is ill-founded and detrimental. The purpose of this commentary is to offer a more supportable perspective on the value of real data examples, to suggest greater restraint in what we ask of examples and what we conclude on their basis, and to offer guidance on using examples effectively for the purposes to which they are well suited.

1. THE BENEFITS OF REAL DATA EXAMPLES

Real data examples offer many benefits (Table 1). First, in the past, to a large extent in textbooks as opposed to journal articles, real data examples served a key role in actually illustrating the computational steps involved in conducting certain statistical tests. This can be invaluable as a pedagogical tool for students and works well with relatively simple calculations. Hence, many statistics books illustrate the calculation of a χ^2 statistic for comparing observed to expected frequencies in small tables. Although this pedagogical tool is enormously valuable, its utility breaks down in complex situations where data sets are necessarily large, cannot be easily summarized by sufficient statistics in simple tables, and require difficult, often iterative, calculations that the reader will not be able to implement with paper and pencil. Hence, the value of examples to illustrate the mechanics of calculations in modern peer-reviewed articles involving genomic techniques is limited.

Second, in the process of applying a newly proposed method to real data, methodologists often find "bugs" in the logic of their proposed method or in their software for implementing it or identify complexities endemic to real data that warrant being addressed via some extension or modification of the method. Thus, the initial application can serve as a useful first field test of the method. Knowing that the method has been applied to real data at least once indicates that the application is practically feasible.

Third, an example can serve another pedagogical purpose, namely, conveying the concept or rationale of the proposed method and illustrating how the results obtained after applying the method can be interpreted. Such uses of examples not only clarify but also can make for more interesting reading. Thankfully, such purposes can be served by any example real data set or for that matter even by a simulated data set. The data set need not be previously unpublished, especially interesting, or yield any particular result.

Fourth, a real data example can provide the author with a vehicle through which to tell a story about why and how the new method should be used. Such storytelling has been shown to help people comprehend and especially retain new ideas more effectively [1].

*Corresponding author.

Table 1. Uses and benefits of real data examples

Use or Benefit of Real Data Example	Comment
Illustrate Computations	Largely passé or inapplicable in peer-reviewed articles of methods for modern problems in genomics
Debug or field test new methods; show that software exists and actually runs in a realistic amount of time.	Can be done with any example real data set. The data set need not be previously unpublished, especially interesting, or yield any particular result.
Illustrate the concept of the method and how results can be interesting.	Can be done with any example real data set or for that matter a simulated data set. The data set need not be previously unpublished, especially interesting, or yield any particular result.
Inspire reader to use new method by serving as an exciting testimonial or case report of the value of the method.	This can be beneficial in promoting use but is tantamount to salesmanship rather than edification.

Finally, real data examples, when they yield biological findings that appear to be new, important, and exciting, can inspire readers to want to use the technique. In our experience, this is a powerful form of inspiration. An attention-getting paper in a premier journal that claims to have an exciting biological finding produced by a new method often initiates a flurry of calls to statistical geneticists by applied scientists wanting help implementing the new catholicon. Although inspiring applied scientists to use new and valuable techniques is meritorious, as we shall discuss below, the increasing demand for *inspiring* examples comes at a price. In our opinion, the price is too steep.

2. THE DISADVANTAGES OF REQUIRING EXAMPLES

Although examples have clear advantages, we believe that making them *de rigueur* or expecting especially exciting ones that produce novel findings has detriments that have gone largely unrecognized. The first detriment is that of promoting the inclusion of extraneous information. That is, in some cases, examples are included because the methodologist knows they are expected and yet they add no additional information or insight to the paper. In many cases, methodologist authors have proven the conclusion of the paper by mathematical proof or simulation study. For some of these cases, it is straightforward to apply the method to real data or it has been demonstrated with simulated data while evaluating the method. Nevertheless, authors may decide to include applications to real data because it is explicitly required or consensus exists that it will strengthen the appeal of the publication. This kind of information regarding real data applications sometimes does not convey any critical information. Furthermore, although the example may not be detrimental, removing it would not affect the fundamental information and logic of the paper [2, 3]. Even if we removed the datasets within the publications by Gauderman et. al, and Jonasdottir et. al., the concept, idea, and logic would still be sound within the publications. Nevertheless, we acknowledge that "Some methods papers are so rarefied that they are virtually inaccessible to the experimentalists who would benefit from the methods. Having the methodologists demonstrate the method in the context of real data forces the methodologist to be more inclusive in their target readership (Copenhaver, personal communication, 12/4/09)."

By way of illustration, consider a hypothetical situation in which a methodologist develops a new method for haplotype phase inference. A good example data set might be one in which phase is known unequivocally so that the accuracy of the new haplotype inference method can be compared to the known phase data. However, if such a data set were unavailable and the methodologist merely applied the new algorithm to a phase-unknown data set, it is unclear what is really learned. Such an exercise would offer no information about the accuracy of the haplotype inference procedure.

A second problem posed by the demand for real data examples, particularly ones that involve previously unpublished data on hot topics and yielding hot findings as "proofs of principles," is that they set unreasonable standards for the methodologist trying to offer his or her work. By definition, hot, new findings are rare. Even the biologist studying at the forefront of some discipline will be very lucky to obtain findings that the scientific community will consider to be of marked interest. Hence, even if a methodologist has developed an outstanding method, it may take many data sets before an analysis yields a result that the scientific community would find truly surprising, interesting, or worthy of special note. Thus, most methodologists will be unable to obtain such examples. Yet, this in no way makes their methodologic development less noteworthy. Moreover, if they were fortunate enough to obtain an exciting new biological finding in a data set, it is likely that this would be done in collaboration with a biologist who "owns" the data set and would likely and understandably wish to publish this first as their primary paper emphasizing the biology and not the methodology. This leads to a situation in which the methodologist must either wait to publish the data set that is no longer brand new and "sexy" or a situation in which a paper is developed that tries to emphasize both the biological findings and the new method that produced them.

We have taken to calling these "*I have an approach*" papers. In our experience, such papers often do not do justice either to the biological phenomenon under inquiry or to the statistical method being developed. Although they are good sales pieces, such papers usually lack rigor in both statistical epistemology[4] and treatment of the biological problem.

A closely related disadvantage of requiring real data examples is that it may deter the creative and forward-looking methodologist from proactively working to develop and publish a new method for analyzing a type of data that is clearly on the horizon, but for which actual data are not yet available. Case in point: whole genome sequencing data 5 years ago.

With respect to developing methods for types of data which may not yet be available, an anonymous reviewer provided us with this very interesting case in which doing so would be challenging. Consider data emerging from lipidomics experiments. Treatments are applied to experimental units (perhaps mutation and wild-type comparisons in plants) and the concentrations of many lipids quantified. Analysis of the resulting data might seem straightforward a priori, yet a complexity of such data is that many zeros can be present. These zeros cannot be ignored and may be present because: (a) The lipid is simply not present in the organism; (b) The lipid is present below detection in some samples; or (c) The reaction that catalyzed the lipid was blocked by the treatment. Reason *c* is extremely important, reason *b* may necessitate some type of imputation below the threshold of detection (which itself may have to be estimated), and reason *a* may simply require omitting that particular lipid from further consideration. Anticipating this issue prior to seeing actual lipidomic data might have been difficult for all but an experienced biochemist. Therefore, it is difficult to imagine statistical methodologists developing new methods which would accommodate this feature of data in advance of seeing the actual data. This point is well taken and certainly any method developed will eventually need to be put to the test in real data, else its validity or lack thereof is of no interest in any case. Thus, ceteris paribus, with new types of data, methodologic papers which can include a real dataset are preferable. And yet, we should not make the perfect the enemy of the good. A method well-conceived in advance of access to a real dataset may itself be useful and, even if it requires extension or modification to accommodate a nuance of the real data, it may nevertheless provide a basis upon which further methodological work can be built.

Finally, a disadvantage of requiring examples that seem to offer proof of principles via their findings is that this promotes what one might call a weak form of fraud among methodologists. Specifically, methodologists who have a new technique that they believe to be better for a particular type of research may analyze a real data set, find that in one instance their method does not seem to offer any advantage over an existing approach, and then move on to another data set, repeating the process until a data set is found in which the new method does indeed yield apparently superior findings. The methodologist then publishes only the example that yielded the apparently better findings. This is promoted by the demand for examples as proof of principle rather than as simply showing how the method proceeds. Because this common use of examples as proof of principle seems to be one of the most problematic aspects of the modern demands for examples, we consider it in greater detail below.

3. INAPPROPRIATE USE OF EXAMPLES: INTERPRETATION AS PROOF OF PRINCIPLE

Spence et al. [5] noted the inappropriateness of relying on a single example as evidence for the value of a statistical method in their theme titled "Willingness to Establish Standards Without the Protections of Rigorous Testing." A key problem with treating real data examples and their findings as proofs of principle for a new method is that this ignores the stochastic component of statistical analysis. That is, statistics typically deals with long-range expectations. Thus, when we say that one method is superior to another in terms of, say, statistical power, we do not mean that the superior method will always yield smaller p values than some less powerful method, only that it will do so more often than not.

To illustrate this point, we conducted a simple simulation with two well-understood tests. The first was the original Haseman-Elston (HE) test of linkage in sibling pairs [6], and the second was the newer weighted HE method of Wang and Elston [7]. It is well established that both of these methods are valid under the null hypothesis, and that under the alternative hypothesis, the newer method is more powerful than the original method. We simulated 1,000 data sets in which a diallelic locus under study had an additive effect explaining 15% of the variance in a quantitative phenotype, the markers were perfectly informative, 1,000 sib pairs were randomly sampled for each of the 1,000 data sets, a background polygenic effect explained 20% of the variance, and minor allele frequency at the locus was 0.2. The resulting p-values are plotted in Figure 1. At the 0.05 α level, the power by the original HE method was 53%. The power by the weighted HE method was 68%, confirming the established power superiority of the newer method. Nevertheless, for over 29% of the data sets, the original less powerful method produced a smaller p-value (i.e., a more significant result) than the more powerful newer method.

What conclusions can we draw from this simulation? These data suggest that had Wang and Elston been required to vet their new method by showing that it produced more significant results than the older method, with realistic data sets, there would have been roughly a 30% chance of finding that the newer method produced less significant results, perhaps leading a naïve reader to conclude that the new method was not really more powerful. Perhaps even more

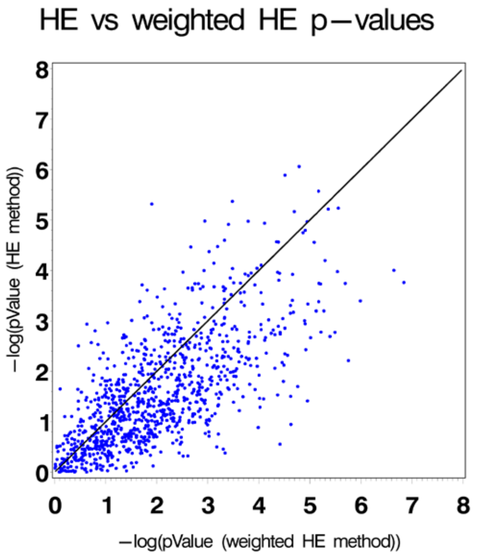

Figure 1. Results (p-values) from analysis of 1,000 simulated data sets analyzed with the original Haseman-Elston (HE) method and a newer weighted Haseman-Elston method.

disturbing, one can imagine that had the newer method in fact been developed earlier, a less mathematically inclined investigator could have subsequently conceived the original HE method, presented it with only an intuitive justification (as has often been done with some of the newer methods in high-dimensional biology [4, 8]), and then, on the basis of a data set that might be chosen 30% of the time, found a real result erroneously offering proof of principle for the power superiority of a less powerful method. Although it is unlikely that this would have occurred in the context of the more statistically mature field of linkage analysis, in the case of, for example, microarray analysis, one can easily imagine such conclusions being drawn.

To illustrate the inappropriate use of examples interpreted as proof of principle, consider the following. Browning [9] offered a "new method for association-based gene mapping that makes powerful use of multilocus data, is computationally efficient, and is straightforward to apply over large genomic regions." As evidence for this claim, Browning offered "analyses of two published data sets that show that this approach can have better power than single-marker tests or sliding-window haplotypic tests." Such a statement vitiates the meaning of power as a statement about long-range frequencies or probabilities.

4. POLICY AND PRACTICE ON REAL DATA EXAMPLES IN MAJOR JOURNALS

In their written policies or guidelines for authors, some journals explicitly state their policy regarding real data examples to encourage, or almost mandate, the inclusion of examples. We examined the policies on real data examples of several major journals that publish statistical methodology papers in the field of applied biological sciences. Excerpts from the aims and scope section of the 9 journals we examined can be found in Supplementary Table 1 online. We also investigated how many journals actually included real data examples among methodological papers published in 2007 from several journals. The proportion of papers with real data examples in each journal can indicate the implicit but actually working policy of editors regarding real data examples.

The journals we examined can be divided into three categories according to their general aims. The first group includes journals whose focus is applied (i.e., not primarily methodological) genetic or genomic research (*American Journal of Human Genetics, Annals of Human Genetics, PLOS Genetics*), but which publish some methodology. The second group consists of journals for which methodological genetic or genomic papers are a staple and major fraction of their output (*Genetic Epidemiology*, *Human Heredity*). The final group consists of statistical journals for methodological application in general biological sciences, which regularly publish some papers related to genomics or genetics (*Biometrics; Biostatistics; Journal of Agricultural, Biological and Environmental Statistics; Statistics in Medicine*). Table 2 summarizes the journals, their general aims, and the policies regarding real data examples.

According to their policy statements, the journals in the third group either encourage or strongly encourage authors to include examples from real data sets. For instance, in its "Information for Authors," *Biometrics* states that the types of papers that will be accepted in its "biometric methodology" section should follow this guideline:

> Regular papers generally focus on the development of new methods and results of use in the biological sciences. These should where possible be made accessible to biologists and other subject-matter scientists by the inclusion of an introductory section outlining the application and scientific objectives on which the new methods focus, with discussion of real data or settings that exemplify the issue being addressed. The journal typically insists on illustration of new methods with real data wherever possible.

Table 3 summarizes the statistical examples provided in the statistical methodology papers in 2007. The information looks at the journal, the number of method papers, and the number of papers with examples as it relates to the three groups of journals in Table 2. As can be seen, within 2007, group 3 had the higher (76% $\sim$ 95%) proportion of papers that included real data set examples.

In the other groups, however, the explicit policy and the actual proportion of real data examples did not always agree. In the first group, *Annals of Human Genetics* had only 28% method papers with real data examples despite their strong policy to include them, whereas all method papers in *PLOS Genetics* included real data examples in 2007

Table 2. Journal policies regarding examples or real data applications

Journal	General Aim	Policy Regarding Examples
American Journal of Human Genetics	Biological research in human genetics and relevant methodological research (Group 1)	Real data application or simulation
Annals of Human Genetics		Real data application is strongly encouraged
PLOS Genetics		No explicit policy
Genetic Epidemiology	Methodological and applied research in genetic epidemiology (Group 2)	No explicit policy
Human Heredity		No explicit policy
Biometrics	Methodological research in biomedical science (Group 3)	Real data application is strongly encouraged
Biostatistics		Real data application is encouraged.
Journal of Agricultural, Biological and Environmental Statistics		Real data application is strongly encouraged
Statistics in Medicine		Real data application is strongly encouraged

Table 3. Journal statistics for the number of papers with examples among statistical methodology papers published in 2007

Journal		Method papers	Papers with examples	
1	*American Journal of Human Genetics*	34	26	(76.5%)
	Annals of Human Genetics	18	5	(27.8%)
	PLOS Genetics	8	8	(100%)
2	*Genetic Epidemiology*	59	35	(59.3%)
	Human Heredity	30	11	(36.7%)
3	*Biometrics*	115	109	(94.8%)
	Biostatistics	58	45	(77.6%)
	Journal of Agricultural, Biological and Environmental Statistics	30	27	(90.0%)
	Statistics in Medicine[*]	139	106	(76.3%)

[*] Among 30 issues of Statistics in Medicine published in 2007, issues 1 ∼ 10 were observed.

without an explicit policy given to the authors. In the second group, these proportions in *Genetic Epidemiology* and *Human Heredity* were quite different (59% and 37%), even though neither had an explicit policy on real data examples and appeared to have very similar scopes for the method papers. Thus, editorial beliefs about the importance of real data examples seem to vary substantially among editors, even for journals with similar policies. Of course, an alternative (or additional) explanation is that perhaps editors often pay little attention to whether there is an example in the paper or not, and it is simply the differences among authors that choose to submit to one sort of journal versus another that results in the different rates.

5. SUMMARY & CONCLUSIONS

Examples can radically enhance a methodologic paper's pedagogical utility and memorableness in many cases. Yet, these goals can be accomplished even with fictitious data sets, just as *Aesop's Fables* effectively offer memorable life lessons even though they are pure fiction. Thus, we strongly advocate the use of examples. However, for the reasons described above, we believe that the demand of some journals, particularly high-profile journals in the genomics arena, for real data examples in general and for examples with "sexy" outcomes is unreasonable and has deleterious consequences. We also suggest that considering single real data examples as proofs of principle is scientifically unsound. Hence, we advocate a relaxation of the expectation for real data examples in general and a dismissal of the idea that the real data examples must have exciting biological findings.

In many articles in this and previous issues of Statistics at Its Interface, authors have used real datasets and enhanced the presentation of their methods. Wu et. al., (2008), proposed a stochastic deletion-insertion algorithm for constructing large-scale linkage maps, and compared it against the seriation, neighbor mapping, and unidirectional growth approach as it modeled a real dataset from the North American Barley Genome Mapping project. Manuscripts from other authors used a simulated data set and a real dataset that demonstrated the usefulness and versatility of the proposed method s illustrated in: Liu et. al. (Controlling Population Structure), Huang et. al. (False-Negative-Rates), Ye et. al. (Clustered Optimal ROC Curve Method), and Li et. al. (2010) (Identify Pathway Regulations in eQTL Mapping) and Tiwari, et. al., (Accurate and Flexible Power Calculations) assessed the performance of their approach against many published findings allowing an assessment of its error rate in practice.

In conclusion, we propose that Editors and Associate Editors make decisions about what is required for papers on a case-by-case basis considering what will best effectively communicate the key messages of the paper in the situation at hand and, of course, what is feasible. The purpose of scientific papers is to communicate something to readers. Therefore, editors and reviewers need to ask themselves "what is important to communicate here and what will it take to effectively communicate it?" as in most of science, we believe that guidelines rather than strict rules work best. In some cases, all that is needed will be a narrative description of a hypothetical example or reference to an example in other published work, in other cases an example showing worked calculations and interpretation may be needed, in still other cases an example involving simulations may be best, and in still other cases application to some real data may be markedly helpful. In each case, we need to ask "What is needed (and feasible) for the paper in question from both a pedagogical perspective and from an epistemological perspective."

ACKNOWLEDGMENTS

We thank Hemant Tiwari for helpful discussion on the topic. Supported in part by NIH grants P30DK056336, U54CA100949, R01ES09912 and T32HL072757. The opinions expressed are those of the authors and not necessarily those of the NIH or any other organization with which the authors are affiliated.

AUTHOR CONTRIBUTIONS

- DBA conceived the paper, directed the analyses, drafted sections, and edited the manuscript.
- AP conducted the simulations and edited the manuscript.
- YJY compiled the information on journal policies and practices, drafted sections of the manuscript, and edited the manuscript.
- KYW drafted sections of the manuscript and edited the manuscript

COMPETING FINANCIAL INTEREST STATEMENT

None of the authors had a competing financial interest to report.

Received 26 May 2010

REFERENCES

[1] Heath, C. and Heath, D. (2007). *Made to Stick: Why Some Ideas Survive and Others Die*. Random House Publishing Group, New York.

[2] Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.*, **31**, 383–395.

[3] Jonasdottir, G., Becker, T., Humphreys, K., and Palmgren, J. (2008). Testing association in the presence of linkage using the GRE and multiple markers. *Genet. Epidemiol.*, **32**, 425–433.

[4] Mehta, T., Tanik, M., and Allison, D. B. (2004). Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.*, **36**(9), 943–947.

[5] Spence, M. A., Greenberg, D. A., Hodge, S. E., and Vieland, V. J. (2003). The emperor's new methods. *Am. J. Hum. Genet.*, **72**, 1084–1087.

[6] Haseman, J. K. and Elston. R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.

[7] Wang, T. and Elston, R. C. (2004). A modified revisited Haseman-Elston method to further improve power. *Hum. Hered.*, **57**, 109–116.

[8] Mehta, T. S., Zakharkin, S. O., Gadbury, G. L., and Allison, D. B. (2006). Epistemological issues in omics and high-dimensional biology: Give the people what they want. *Physiol. Genomics.*, **28**(1), 24–32.

[9] Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.*, **78**(6), 903–913.

K. Y. Williams
Department of Biostatistics
Section on Statistical Genetics
University of Alabama at Birmingham
Birmingham, Alabama
USA

Yun Joo Yoo
Department of Mathematics Education
Seoul National University
Kwanak-ro 1, Kwanak-ku, Seoul
South Korea

Amit Patki
Department of Biostatistics
Section on Statistical Genetics
University of Alabama at Birmingham
Birmingham, Alabama
USA

David B. Allison
Department of Biostatistics
Section on Statistical Genetics

Department of Nutrition Sciences

Clinical Nutrition Research Center
University of Alabama at Birmingham
Birmingham, Alabama

Ryals Public Health Building
Suite 327
1665 University Boulevard
Birmingham, Alabama 35294
USA
Tel: 205-975-9169
Fax: 205-975-2540
E-mail: Dallison@UAB.edu